PR #24682 完整报告

sgl-project/sglang

[diffusion] doc: update ltx2 multi-gpu deployment guide

合并时间：2026-05-08 18:38

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24682

执行摘要

本次 PR 更新了 LTX 系列模型的多 GPU 部署文档，在交互式命令生成器中新增 2/4 GPU 预设，同时将 cookbook 页面重命名为 LTX2 & LTX2.3，添加了 Fast 多 GPU 部署表格和徽标，为 CFG parallel 和 TP 参数提供直观指导。

功能与动机

原文档仅覆盖单 GPU 场景，用户需要了解如何利用多 GPU 加速 LTX 推理。新增的预设帮助用户快速生成正确的 sglang serve 命令，避免手动配置错误。

实现拆解

1. 在 ltx-deployment.jsx 中增加 h200-2gpu 和 h200-4gpu 硬件选项，并实现 getParallelFlags() 函数，根据硬件返回对应的 --num-gpus 和 --enable-cfg-parallel 等参数。
2. 将 LTX.mdx 重命名为 LTX2 & LTX2.3.mdx，更新标题和描述，新增“Fast multi-GPU presets”表格，对比不同 GPU 数量下的推荐参数。
3. 更新 docs.json 中的导航路径，指向新页面。
4. 添加 ltx.svg 徽标，并更新 intro.mdx 中的卡片链接和图片。

docs_new/src/snippets/diffusion/ltx-deployment.jsx

核心交互组件，新增多 GPU 硬件选项和参数生成逻辑

```
const getParallelFlags = () => {
  // 使用对象查找维护并行配置，方便后续新增硬件选项
  const parallelFlagsMap = {
    'h200-2gpu': ` --num-gpus 2 --enable-cfg-parallel`,
    'h200-4gpu': ` --num-gpus 4 --tp-size 2 --enable-cfg-parallel`,
  };
  return parallelFlagsMap[values.hardware] || '';
};

const getDeviceMode = () => {
  // 使用 startsWith 匹配所有 h200-* 前缀，提升可扩展性
  if (values.hardware.startsWith('h200')) {
    return 'resident';
  }
```

```
  // ...
};
```

评论区精华

> gemini-code-assist[bot] 建议：用 startsWith('h200') 替代逐个 === 判断，提升可扩展性；将 if-else 链重构为对象查找 parallelFlagsMap，便于未来添加更多配置。两个建议均被采纳，最终代码按此方式实现。

风险与影响

风险极低，仅涉及文档和前端组件。若命令行参数生成有误可能误导用户，但经过 Review 和实际测试可降低此风险。

关联脉络

本 PR 延续了 SGLang 对 diffusion 模型文档的持续完善，与近期多个 LTX 性能优化（如 FA3 修复 #24688、帧返回优化 #24616）共同提升了 LTX 模型的可用性。